Attorney Docket No.: YOR920030277US1 (8728-634)

U.S. Patent Application:

Title: APPARATUS FOR SCALABLE RELIABLE GROUP COMMUNICATION

Inventors:

Francois Baccelli
83 rue de Paris
Meudon 92190, FRANCE

Augustin Chaintreau
19, rue Saint Didier
75116 Paris, FRANCE

Zhen Liu
37 Roundabend Road
Tarrytown, New York 10591

Anton Riabov
452 Riverside Drive, Apt. 61
New York, New York 10027

Sambit Sahu
551 Kennicut Hill Road
Mahopac, New York 10541

Assignee: International Business Machines Corporation

Date Deposited: November 25, 2003

F. Chau & Associates, LLP
1900 Hempstead Turnpike, Suite 501
East Meadow, NY 11554
Tel: (516) 357-0091
Fax: (516) 357-0092

APPARATUS FOR SCALABLE RELIABLE GROUP COMMUNICATION

BACKGROUND OF THE INVENTION

1. Field of the Invention:

The present invention relates to communications networks, and more particularly to a system and method for reliable data delivery to a group of receivers using overlay distribution tree.

2. Discussion of Related Art:

Reliable delivery of content to a group of receivers has several applications. IP-multicast based solutions have been advocated to address content delivery to groups. However, due to deployment issues and scalability concerns, IP-multicast has not succeeded in providing large-scale reliable group communication.

An example of an IP-multicast network is shown in Figure 1. The network includes an origin node 101 and receiving nodes, e.g., 102 connected through TCP servers, e.g., 103. As is illustrated, each communication comprises multiple connections between, for example, the origin node and a server, and between the server and the receiving node.

While IP-multicast has been examined and matured into a communication mechanism for group-based communication, because of deployment and scalability issues, it is not an attractive solution. More particularly, IP-multicast is not widely deployed in the

Internet in spite of extensive research as well as industrial efforts. Further, the throughput of TCP-based reliable communication decreases as the inverse of log of number of participants if IP-multicast is used. Thus, IP-multicast is difficult to scale.

Therefore, a need exists for a system and method for a scalable overlay distribution tree.

SUMMARY OF THE INVENTION

According to an embodiment of the present invention, a method for group communication over a network of processors comprises determining an overlay spanning tree comprising an origin node and at least one receiving node, and controlling a source communication rate to be less than or equal to a bottleneck rate of the overlay spanning tree.

The method comprises protecting data delivery by link error recovery. The overlay spanning tree comprises a plurality of nodes, wherein the data delivery is reliable such that each node receives the same data.

The method comprises scaling the overlay spanning tree to an arbitrary group size.

The method further comprises determining a maximum throughput of the spanning tree among all possible configurations of the spanning tree given a reduced overlay distribution tree. Determining the overlay spanning tree comprises defining a target bandwidth for the overlay tree given a fully connected overlay distribution graph, constructing a reduced overlay distribution graph by removing an edge from the fully connected overlay distribution graph having a bandwidth less than or equal to the target bandwidth, and constructing an arbitrary spanning tree comprising a root, wherein the root is a source node of a plurality of links in the reduced overlay distribution graph. Determining the overlay spanning tree further comprises performing a triangular improvement to remove a link violating a rate constraint, increasing the target bandwidth upon determining that the overlay spanning tree is constructible, and decreasing the target bandwidth upon determining that the overlay spanning tree is not constructible.

The method comprises joining a new node to the spanning tree. The method further comprises joining the new node to an existing node of the spanning tree upon determining that the existing node has a bandwidth of greater than or equal to an existing rate. The method comprises determining a triangular improvement upon determining that no existing node has a bandwidth greater than or equal to the existing rate, joining the new node at an attachment point having a highest bandwidth among existing nodes of the spanning tree upon determining that the triangular improvement failed, and redetermining the spanning tree upon determining bandwidth less than or equal to a minimum threshold.

The method comprises redetermining the spanning tree upon determining that an existing node has left the spanning tree. The method comprises determining orphaned child nodes of the existing node that has left the spanning tree, and performing a join for each orphaned child node.

According to an embodiment of the present invention, a program storage device is provided readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for group communication over a network of processors. The method comprising determining an overlay spanning tree comprising an origin node and at least one receiving node, and controlling a source communication rate to be less than or equal to a bottleneck rate of the overlay spanning tree.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the present invention will be described below in more detail, with reference to the accompanying drawings:

Figure 1 is a diagram of an IP-multicast communications group;

Figure 2 is a diagram of a system according to an embodiment of the present invention;

Figure 3 is a diagram of a communications group according to an embodiment of the present invention;

Figure 4 is a flow chart of a method according to an embodiment of the present invention; and

Figure 5 is a flow chart of a method according to an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Reliable delivery of content to a group of receivers has several applications. For example, collaborative applications delivering content to a set of populations without consuming undesirable quantities of network resources and without disrupting other existing forms of communications.

The present invention is both scalable and efficient in supporting reliable content delivery to a set of populations. An overlay distribution tree is built where receivers are the nodes in the tree. Using rate control at an origin node and buffer management at the application layer at each node, TCP congestion control is used locally at each node to reliably transfer the content.

It is to be understood that the present invention may be implemented in various forms of hardware, software, firmware, special purpose processors, or a combination thereof. In one embodiment, the present invention may be implemented in software as an application program tangibly embodied on a program storage device. The application program may be uploaded to, and executed by, a machine comprising any suitable architecture.

Referring to Figure 2, according to an embodiment of the present invention, a computer system 201 for implementing the present invention can comprise, *inter alia*, a central processing unit (CPU) 202, a memory 203, and an input/output (I/O) interface 204. The computer system 201 is generally coupled through the I/O interface 204 to a display 205 and various input devices 206 such as a mouse and keyboard. The support circuits can include circuits such as cache, power supplies, clock circuits, and a communications bus. The memory 203 can include random access memory (RAM), read only memory (ROM), disk drive, tape drive, etc., or a combination thereof. The present invention can be implemented as a routine 207 that is stored in memory 203 and executed by the CPU 202 to process the signal from the signal source 208. As such, the computer system 201 is a general-purpose computer system that becomes a specific purpose computer system when executing the routine 207 of the present invention.

The computer platform 201 also includes an operating system and microinstruction code. The various processes and functions described herein may either be part of the microinstruction code or part of the application program (or a combination thereof), which is executed via the operating system. In addition, various other peripheral devices may be connected to the computer platform such as an additional data storage device and a printing device.

It is to be further understood that, because some of the constituent system components and method steps depicted in the accompanying figures may be implemented in software, the actual connections between the system components (or the process steps) may differ depending upon the manner in which the present invention is programmed. Given the teachings of the present invention provided herein, one of ordinary skill in the related art will be able to contemplate these and similar implementations or configurations of the present invention.

According to an embodiment of the present invention, the TCP stack and buffering at the application layer are used to support reliable group communication on an overlay distribution tree. No special support is needed from the routers - the overlay distribution tree can be deployed in the Internet with the existing protocols in the routers. The overlay distribution tree can be implemented for any group size, the method is scalable.

Referring to Figure 3, a congestion control protocol of TCP is implemented in a hop-by-hop manner by using the overlay distribution tree to forward contents. A hop denotes the connection path 301 between any two participants, e.g., an origin node 302 and a receiving node 303. Thus, if an origin node throttles the sending rate at or below a bottleneck link bandwidth in the overlay tree, it can scale to any arbitrary group size. That is, by preventing bottlenecks through rate control, network reliability is maintained for different scales.

Different methods can be used to generate the overlay distribution tree. For example, according to an embodiment of the present invention, the distribution tree achieves a high throughput when TCP congestion control is used between any link for reliably transferring data to a set of participants.

Referring to Figure 4, assuming that each node has knowledge about its access link bandwidth, and that the end-to-end TCP bandwidth is known between any two nodes, a TCP connection is opened between two nodes 401. Let this be denoted as tcp(i,j) between any two nodes i and j. Let a(i) be the access link bandwidth from a node i.

Starting with a fully connected graph, the maxB is determined 402. The maxB is the maximum link bandwidth where link bandwidth on edge (i,j) is given by min {a(i), a(j), tcp(i,j)}.

Note that the optimal group bandwidth will be between 0 and maxB. Thus,

minB =0;

maxB = maximum link bandwidth;

A target bandwidth is defined as 403:

targetB = (minB + maxB)/2

where the result can be rounded, preferably down.

The edges that have less link bandwidth than targetB are removed 404. Here link bandwidth is given by min{a(i)/f(i), tcp(i,j), a(j)/f(j)} where f(i) is the outgoing edges from node i. It is determined whether the resulting graph is a disconnected graph 405. If the graph is disconnected, set maxB = targetB 407.

A spanning tree is determined from the reduced graph 406 upon determining that the graph is not disconnected. The spanning tree is constructed having a desirable throughput, e.g., a spanning tree having a maximum throughput among all possible configurations of the spanning tree. A violation index V(i) is determined for each node i which is defined as target B/a(i) * f(i) 501. If it is determined that a node has a positive violation index 502, set minB = targetB 503. Upon determining that the node has a positive violation index, the target bandwidth is redefined 503. Upon determining that no node has a positive violation index, the node in the spanning tree that has the largest violation index is determined 504. A triangular improvement is applied to reduce the violation index of the node 505. If the violation index of the node is determined to not have been reduced set maxB = targetB, and the target bandwidth is redefined 506. Upon determining that the violation index of the node has been reduced, the violation index of a next node is determined 501.

The two nodes that have the next largest violation index are determined, such that there exists edges to these two nodes from this largest violating node. The existing link is replaced and the links that have the next largest violation link are added. A depth first algorithm could be used to find the existence of such two nodes. This process is triangular improvement - in which the violation index of the node under consideration is reduced by one.

The spanning tree is reduced until a condition is reached wherein all nodes are non-violating. If such a tree does not exist, set maxB = targetB, and go to block 403.

Once the spanning tree has been reduced to where all nodes are non-violating, set minB = targetB, and go to block 403.

The binary search on minB and maxB determines the overlay that results in the maximum group throughput. Let the maximum rate for which there is a non-violating spanning tree be denoted as rateB.

The sending rate of origin node is set to rateB. Use the resulting overlay tree to send data using the hop-by-hop TCP congestion control.

According to an embodiment of the present invention, a method for constructing a scalable spanning tree to group communication is within ½ of an optimal solution and shown to scale to any arbitrary group size. Note that the above constructed overlay handles a given set of nodes. The proposed solution can be adapted to handle leave and joining of nodes in the communication tree.

Join/leave operations can be performed according a desired protocol. For example, if the new participating node has a link to a node in the existing overlay with sufficient bandwidth, the new node is attached to that available node. If not, the new node is attached to any arbitrary link and the method described with respect to Figure 4 is applied. For nodes leaving the overlay distribution tree, e.g., the communications group, these nodes are treated as a set of nodes joining. Consider that node i is leaving from the overlay. Let the children of node i be denoted as Child(i). A new join procedure is followed for each node j in the set Child(i). Methods for handling join/leave can be improved to have substantially similar performance as a solution for the static case, e.g., no join/leave.

Having described embodiments for a system and method for reliable content delivery to a set of receivers, it is noted that modifications and variations can be made by persons skilled in the art in light of the above teachings. It is therefore to be understood that changes may be made in the particular embodiments of the invention disclosed which are within the scope and spirit of the invention as defined by the appended claims. Having thus described the invention with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.